Prasham Shah

San Diego, CA | 626-734-8075 | prs008@ucsd.edu | linkedin.com/in/prashamshahh | www.prasham.ai

Education

University of California, San Diego **Sep 2023 – Jun 2027**
B.S. in Computer Science | GPA: 3.7 *San Diego, CA*

Experience

Founder & Software Engineer **Aug 2025 – Present**
Novalo *San Diego, CA*

- Built and operate a **multi-tenant AI productivity platform** serving **100+ weekly users**, supporting document ingestion, task workflows, and real-time assistant features.
- Designed backend services for **retrieval**, **background jobs**, and **model routing**, balancing latency and cost across heterogeneous AI workloads while maintaining **sub-2s response times**.
- Integrated **structured outputs**, **scheduled workflows**, and external tooling for notifications and automation, improving production reliability and enabling evaluation-driven iteration.

Software Engineering Research Assistant **Jun 2025 – Present**
UCSD DigiHealth Lab *San Diego, CA*

- Developed **secure full-stack infrastructure** for clinical AI research with authenticated workflows, validation layers, and audit-conscious handling of sensitive data.
- Deployed and performance-tuned backend services for **LLM-based triage tools** on dedicated servers, achieving **sub-1.2s median latency** under concurrent load.
- Served fine-tuned open-source models via **vLLM**, optimizing inference pipelines to reduce compute cost by **80%** versus API-based alternatives.

Software Engineer **Sep 2024 – Sep 2025**
BuyFineDiamonds (BFD) *London, UK / Remote*

- Designed and operated backend infrastructure for a luxury e-commerce platform with **100K+ SKUs**, building APIs for **search**, **ranking**, and **personalization**.
- Engineered **FastAPI** services with **async I/O**, **Redis caching**, and optimized query paths, maintaining **sub-300ms P95 latency** under production traffic.
- Reduced end-to-end request latency by **65%** through caching and database optimization, decreasing **PostgreSQL** load by **40%** and improving system reliability.

Projects

OpenCL Accel. for Financial Time Series, CNN Workloads | C++, Python **Oct 2025 – Mar 2026**

- Built **GPU-accelerated OpenCL pipelines** for financial time-series feature generation and CNN-style workloads, parallelizing rolling-window statistics, convolution, and softmax across large batched datasets.
- Optimized kernels using **tiling**, **local memory**, **coarsening**, and reduced host-device transfer overhead, achieving up to **10x speedups** over CPU baselines on matrix-heavy workloads.
- Analyzed **arithmetic intensity**, **memory traffic**, and kernel bottlenecks using FLOP and memory-access models to guide performance tuning and architecture-aware optimization.

Real-Time Analytics Platform | JavaScript, Node.js, PostgreSQL, Nginx, Docker **Jan 2026 – Mar 2026**

- Built an end-to-end analytics platform from scratch with a **client-side collector**, backend ingestion and reporting services, and dashboards for **traffic**, **event**, and **behavioral** analysis.
- Instrumented a live website to capture **page views**, **sessions**, **custom events**, **performance signals**, and request metadata, then processed and stored the data through a production-style backend pipeline.
- Deployed the system on real infrastructure with **Nginx**, domain routing, and server logging, demonstrating full ownership from collection and transport through storage and visualization.

Technical Skills

Languages: Python, C, C++, TypeScript, JavaScript, SQL, OpenCL
Frameworks & Backend: FastAPI, Node.js, React, Next.js, REST APIs, WebSockets
Systems & Infrastructure: Docker, Nginx, Redis, PostgreSQL, Supabase, AWS, Linux, caching, deployment, distributed systems
AI / ML Systems: vLLM, RAG pipelines, model routing, vector search, fine-tuning, LLM application workflows
Core Areas: Data Structures and Algorithms, System Design, Parallel Computing, Server-side Dev